Research Statement

Ryan Giordano

Many researchers would be concerned if they learned that some core conclusion of their statistical analysis—such as the sign or statistical significance of some key effect—could be overturned by removing a small fraction, say 0.1%, of their data. Such non-robustness would be particularly concerning if the data was not actually drawn randomly from precisely the population of interest, or if the model may have been misspecified—circumstances that often obtain in the social sciences, for example. Nevertheless, analysts do not routinely check whether ablation of such a small set could overturn their results, in part because the number of possible subsets containing 0.1% of the data points is combinatorially large.

In Broderick et al. [2020], I identify problematic subsets of the data by forming a *linear approximation* to how a wide class of statistical estimators depend on their datasets. The key idea is that, although there are a very large number of subsets containing 0.1% of the data points, none of them are very different from the original dataset, so the linear approximation will be accurate for precisely the sort of small proportions that are worrying for robustness. I provide finite-sample accuracy bounds for the approximation, an `R` package to compute the approximation automatically using automatic differentiation [Baydin et al., 2017, Maclaurin et al., 2017, Giordano, 2020], and we show that the approximation is capable of detecting meaningful non-robustness in published econometrics studies. For example, in a study of microcredit in Mexico [Angelucci et al., 2015], we find that, by removing just 15 households out of 16,561 studied (a change of less than 0.1%), the estimated effect of microcredit changes from negative and statistically insignificant to positive and statistically significant.

In particular, I form an approximation for estimators that can be expressed as the root of a smooth estimating equation—a broad class that includes maximum likelihood (MLE), ordinary least squares (OLS), instrumental variable (IV), generalized method of moments (GMM), and variational Bayes (VB) estimators, among others. The approximation can be computed quickly and automatically using automatic differentiation [Baydin et al., 2017, Maclaurin et al., 2017], and I provide an `R` package to do so (`zaminfluence` [Giordano, 2020]). By ablating the subset identified by the approximation re-computing the estimator only once, one can form a lower bound on the sensitivity since, at worst, the approximation identified a sub-optimal subset.

An approximation is only as useful as its accuracy. Using techniques from Giordano et al. [2019a], I provide finite-sample accuracy bounds for the approximation in terms of the set complexity of the estimating equation and its derivatives; for OLS and IV, our error bounds are exactly computable using only the regression output. I show that the approximation's relative error is small when the proportion of data removed, even if the total number of data points removed is very large.

By studying properties of the approximation, I show that our metric captures a type of non-robustness that is qualitatively different from classical robustness. Non-robustness to the ablation of small datasets is driven by low signal-to-noise ratio in the inference problem, is not reflected in standard errors, does not disappear asymptotically, and is not a product of misspecification.

To show that we can detect meaningful non-robustness in real datasets, my co-authors and I applied our methods to a number of published studies in econometrics. Though some results were robust to the ablation of small subsets of the data, others were not. For example, in a study of microcredit in Mexico [Angelucci et al., 2015], we find that, by removing just 15 households out of 16,561 studied (a change of less than 0.1%), the estimated effect of microcredit changes from negative and statistically insignificant to positive and statistically significant.

The core idea underlying the above work underlies all of my current research, and it is this: many tasks in data science are computationally difficult because they require re-computation of a statistical estimator on inputs that are, in some sense, "near" the original inputs, and Taylor series approximations can provide fast and accurate approximations to expensive re-computation. In my work, I show that essentially the same ideas can provide fast approximations to cross validation, the bootstrap, Bayesian prior sensitivity, and even posterior covariance estimation for mean field variational Bayes. For the remainder of this essay, I will discuss each of these applications in turn, emphasizing the ways in which I update classical results with intuitive, relevant theory and easy-to-use computational tools.

Approximate cross validation.

The error or variability of machine learning algorithms is often assessed by repeatedly re-fitting a model with different weighted versions of the observed data; cross-validation (CV) can be thought of as a particularly popular example of this technique. In Giordano et al. [2019b], I use a linear approximation to the dependence of the fitting procedure on the weights, producing results that can be faster than repeated re-fitting by an order of magnitude. I provide explicit finite-sample error bounds for the approximation in terms of a small number of simple, verifiable assumptions. My results apply whether the weights and data are stochastic or deterministic, and so can be used as a tool for proving the accuracy of the infinitesimal jackknife on a wide variety of problems. As a corollary, I state mild regularity conditions under which the approximation consistently estimates true leave-k-out cross-validation for any fixed k. I demonstrate the accuracy of the approximation on a range of simulated and real datasets, including an unsupervised clustering problem from genomics [Luan and Li, 2003, Shoemaker et al., 2015].

Approximately bootstrapping Bayesian posterior means.

The frequentist (i.e., sampling) variance of Bayesian posterior expectations differs in general from the posterior variance even for large datasets, particularly when the model is misspecified or contains many latent variables [Kleijn and van der Vaart, 2006]. Knowing the frequentist variance of a posterior expectation can be useful even to a committed Bayesian, particularly when the data is known to arise from random sampling and there is a possibility of model misspecification [Waddell et al., 2002]. However, the principal existing approach for computing the frequentist variability from MCMC procedures is the bootstrap, which can be extremely computationally intensive due to the need to run hundreds of extra MCMC procedures [Huggins and Miller, 2019].

In [Giordano and Broderick, 2020a,b], I propose an efficient alternative to bootstrapping an MCMC procedure which is based on the influence function from sensitivity analysis. Using results from [Giordano et al., 2018a, 2019b], I show that the influence function for posterior expectations can be easily computed from the posterior samples of a single MCMC procedure and consistently estimates the bootstrap variance. I demonstrate the accuracy and computational benefits of the influence function variance estimates on array of experiments including an election forecasting model [Gelman and Heidemanns, 2020], the Cormack-Jolly-Seber model from ecology [Kéry and Schaub, 2011], and a large collection of models and datasets from the social sciences [Gelman and Hill, 2006].

Bayesian prior sensitivity.

Prior sensitivity for discrete Bayesian nonparametrics. A central question in many probabilistic clustering problems is how many distinct clusters are present in a particular dataset. A Bayesian nonparametric (BNP) model addresses this question by placing a generative process on cluster assignment, making the number of distinct clusters present amenable to Bayesian inference. However, like all Bayesian approaches, BNP requires the specification of a prior, and this prior may favor a greater or lesser number of distinct clusters.

In [Giordano et al., 2018c], I derive prior sensitivity measures for a truncated variational Bayes approximation using ideas from [Gustafson, 1996, Giordano et al., 2018a]. Unlike previous work on local Bayesian sensitivity for BNP [Basu, 2000], I pay special attention to the ability of the sensitivity measures to *extrapolate* to different priors, rather than treating the sensitivity as a measure of robustness *per se*. In work currently in progress [Liu et al., 2020], my co-author and I apply the approximation from [Giordano et al., 2018c] to an unsupervised clustering problem on a human genome dataset [Huang et al., 2011, Raj et al.,

2014], demonstrating that the approximate is accurate, orders of magnitude faster than re-fitting, and capable of detecting meaningful prior sensitivity.

Prior sensitivity for Markov Chain Monte Carlo. MCMC is arguably the most commonly used computational tool to estimate Bayesian posteriors, which is made still easier by modern black-box MCMC tools such as `Stan` [Carpenter et al., 2017, Stan Development Team, 2020]. However, a single run of MCMC typically remains time-consuming, and systematically exploring alternative prior parameterizations by re-running MCMC would be computationally prohibitive for all but the simplest models.

My software package, `rstansensitivity`, [Giordano, 2018, Giordano et al., 2018b], takes advantage of the automatic differentiation capacities of `Stan` [Carpenter et al., 2015] together with a classical result from Bayesian robustness [Gustafson, 1996, Basu et al., 1996, Giordano et al., 2018a] to provide automatic hyperparameter sensitivity for generic `Stan` models from only a single MCMC run. I demonstrate the speed and utility of the package in detecting excess prior sensitivity, particularly in a social sciences model taken from Gelman and Hill [2006, Chapter 13.5].

Uncertainty propagation in mean-field variational Bayes.

Mean-field Variational Bayes (MFVB) is an approximate Bayesian posterior inference technique that is increasingly popular due to its fast runtimes on large-scale scientific data sets (e.g., Raj et al. [2014], Kucukelbir et al. [2017], Regier et al. [2019]). However, even when MFVB provides accurate posterior means for certain parameters, it often mis-estimates variances and covariances [Wang and Titterington, 2004, Turner and Sahani, 2011] due to its inability to propagate Bayesian uncertainty between statistical parameters.

In Giordano et al. [2015, 2018a], I derive a simple formula for the effect of infinitesimal model perturbations on MFVB posterior means, thus providing improved covariance estimates and greatly expanding the practical usefulness of MFVB posterior approximations. The estimates for MFVB posterior covariances rely on a result from the classical Bayesian robustness literature that relates derivatives of posterior expectations to posterior covariances and includes the Laplace approximation as a special case. In the experiments, I demonstrate that my methods are simple, general, and fast, providing accurate posterior uncertainty estimates and robustness measures with runtimes that can be an order of magnitude faster than MCMC, including models from ecology [Kéry and Schaub, 2011], the social sciences [Gelman and Hill, 2006], and on a massive internet advertising dataset [Criteo Labs, 2014].

Selected Future work

My research is ideally driven by the needs of my scientific and industry collaborators, and so I expect my future work will be determined to a large part by my colleagues. However, I will now discuss a few directions that I find promising and interesting, and which I believe could be applicable to a diverse set of problems.

The higher-order infinitesimal jackknife for the bootstrap. In the preprint Giordano et al. [2019a], I extend Giordano et al. [2019b] to higher-order Taylor series approximations, providing a family of estimators which I collectively call the higher-order infinitesimal jackknife (HOIJ). In addition to providing higher-quality approximations to CV and extending the results to k-fold CV, the higher-order approach promises to provide a scalable alternative to the bootstrap, a procedure that estimates frequentist variability by repeatedly re-evaluating a model at datasets drawn with replacement from the observed data. The bootstrap is known to enjoy higher-order accuracy in certain circumstances Hall [2013], and the HOIJ can approach the bootstrap at a rate faster than the bootstrap approaches the

truth. The HOIJ thus promises to make bootstrap inference available to models which are differentiable but too expensive to re-evaluate (e.g. simulation-based models [Baker et al., 2019, Section 2.6]), but also to allow efficient bootstrap-after-bootstrap procedures which that are currently out of reach for all but the simplest statistics [Efron and Tibshirani, 1994].

Scaling sensitivity measures. Sensitivity analysis typically avoids the expense of re-fitting a model, but incurs the expense of solving a large linear system. Thus, extending the benefits of the sensitivity analysis to increasingly large scientific problems requires developing methods to efficiently solve correspondingly large linear systems. Stochastic second-order methods are currently an active research topic in optimization [Agarwal et al., 2017, Berahas et al., 2020], and methods developed therein should apply directly to sensitivity analysis.

Partitioned Bayesian inference. The ideas of [Giordano et al., 2018a] can be naturally extended to approximately propagate uncertainty among separately estimated components of an inference problem. For example, astronomical catalogs are customarily produced with MFVB-like algorithms [Lang et al., 2016, Regier et al., 2019], which take inputs such as the sky background and optical point spread function as fixed inputs, though these quantities are themselves inferred with uncertainty. Viewing all the separate inference procedures as a sequential quasi-MFVB objective, one could directly apply the techniques of [Giordano et al., 2018a] to propagate the uncertainty from the modeling inputs to the astronomical catalog's uncertainty.

References

Agarwal, N., Bullins, B., and Hazan, E. (2017). Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1):4148–4187.

Angelucci, M., Karlan, D., and Zinman, J. (2015). Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82.

Baker, N., Alexander, F., Bremer, T., Hagberg, A., Kevrekidis, Y., Najm, H., Parashar, M., Patra, A., Sethian, J., Wild, S., Wilcox, K., and Lee, S. (2019). Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence. Technical report, USDOE Office of Science (SC), Washington, DC (United States).

Basu, S. (2000). Bayesian robustness and Bayesian nonparametrics. In Insua, D. R. and Ruggeri, F., editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media.

Basu, S., Jammalamadaka, S. R., and Liu, W. (1996). Local posterior robustness with parametric priors: Maximum and average sensitivity. In *Maximum Entropy and Bayesian Methods*, pages 97–106. Springer.

Baydin, A., Pearlmutter, B., Radul, A., and Siskind, J. (2017). Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18(153):1–153.

Berahas, A., Bollapragada, R., and Nocedal, J. (2020). An investigation of Newton-sketch and subsampled Newton methods. *Optimization Methods and Software*, pages 1–20.

Broderick, T., Giordano, R., and Meager, R. (2020). An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *arXiv preprint arXiv:2011.14999*. Authors listed alphabetically. Giordano, R. and Meager, R. are equal contribution primary authors.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Carpenter, B., Hoffman, M., Brubaker, M., Lee, D., Li, P., and Betancourt, M. (2015). The Stan math library: Reverse-mode automatic differentiation in C++. *arXiv preprint arXiv:1509.07164*.

Criteo Labs (2014). Criteo conversion logs dataset. Downloaded on July 27th, 2017.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. CRC press.

Gelman, A. and Heidemanns, M. (2020). The Economist: Forecasting the US elections. Data and model accessed Oct., 2020.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Giordano, R. (2018). StanSensitivity: Automated hyperparameter sensitivity for Stan models. GitHub repository `https://github.com/rgiordan/StanSensitivity`.

Giordano, R. (2020). Zaminfluence. GitHub repository `https://github.com/rgiordan/zaminfluence`.

Giordano, R. and Broderick, T. (2020a). The Bayesian infinitesimal jackknife for variance. *In preparation.*

Giordano, R. and Broderick, T. (2020b). Effortless frequentist covariances of posterior expectations in Stan. Presentation at Stancon 2020 `https://tinyurl.com/y2e2ucp3`.

Giordano, R., Broderick, T., and Jordan, M. (2018a). Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029.

Giordano, R., Broderick, T., and Jordan, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449.

Giordano, R., Broderick, T., and Jordan, M. I. (2018b). Automatic robustness measures in Stan. Presentation at Stancon 2018 `https://tinyurl.com/yyqwpowc`.

Giordano, R., Jordan, M. I., and Broderick, T. (2019a). A higher-order Swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*.

Giordano, R., Liu, R., Jordan, M. I., and Broderick, T. (2018c). Evaluating sensitivity to the stick breaking prior in Bayesian nonparametrics. *arXiv preprint arXiv:1810.06587*.

Giordano, R., Stephenson, W., Liu, R., Jordan, M. I., and Broderick, T. (2019b). A Swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147.

Gustafson, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195.

Hall, P. (2013). *The Bootstrap and Edgeworth Expansion*. Springer Science & Business Media.

Huang, L., Jakobsson, M., Pemberton, T., Ibrahim, M., Nyambo, T., Omar, S., Pritchard, J., Tishkoff, S., and Rosenberg, N. (2011). Haplotype variation and genotype imputation in African populations. *Genetic epidemiology*, 35(8):766–780.

Huggins, J. and Miller, J. (2019). Using bagged posteriors for robust inference and model criticism. *arXiv preprint arXiv:1912.07104*.

Kéry, M. and Schaub, M. (2011). *Bayesian population analysis using WinBUGS: A hierarchical perspective*. Academic Press.

Kleijn, B. and van der Vaart, A. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2):837–877.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

Lang, D., Hogg, D., and Mykytyn, D. (2016). The Tractor: Probabilistic astronomical source detection and measurement. *ascl*, pages ascl–1604.

Liu, R., Giordano, R., and Broderick, T. (2020). Evaluating sensitivity to the stick breaking prior in Bayesian nonparametrics. Presentation at BAYESM:O 2020 `https://tinyurl.com/y2qgofgl`.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, 19(4):474–482.

Maclaurin, D., Duvenaud, D., and Johnson, M. (2017). autograd. GitHub repository `https://github.com/HIPS/autograd`.

Raj, A., Stephens, M., and Pritchard, J. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589.

Regier, J., Fischer, K., Pamnany, K., Noack, A., Revels, J., Lam, M., Howard, S., Giordano, R., Schlegel, D., and McAuliffe, J. (2019). Cataloging the visible universe through Bayesian inference in Julia at petascale. *Journal of Parallel and Distributed Computing*, 127:89–104.

Shoemaker, J. E., Fukuyama, S., Eisfeld, A. J., Zhao, D., Kawakami, E., Sakabe, S., Maemura, T., Gorai, T., Katsura, H., Muramoto, Y., Watanabe, S., Watanabe, T., Fuji, K., Matsuoka, Y., Kitano, H., and Kawaoka, Y. (2015). An ultrasensitive mechanism regulates influenza virus-induced inflammation. *PLoS Pathogens*, 11(6):1–25.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time Series Models*.

Waddell, P., Kishino, H., and Ota, R. (2002). Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. *Genome Informatics*, 13:82–92.

Wang, B. and Titterington, M. (2004). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Workshop on Artificial Intelligence and Statistics*, pages 373–380.